Assembling crop genomes with 2\textsuperscript{nd} and 3\textsuperscript{rd} generation sequencing

Michael Schatz

Oct 8, 2012
Strategies for de novo assemblies of complex crop genomes
The Genome Analysis Center, Norwich Research Park

#ESFCrops / @mike_schatz
Outline

1. Ingredients for a good assembly

2. 2nd Generation Sequencing & Assembly
   1. Sacred Lotus
   2. Raspberry
   3. Wheat

3. 3rd Generation Sequence & Assembly
   1. Parrot
   2. Rice
Assembling a Genome

1. Shear & Sequence DNA

2. Construct assembly graph from overlapping reads

\[...AGCCTAGACCTACA\] \[GGATGCGCGACACGT\] \[GGATGCGCGACACGT\] \[CGCATATCCGGT...\]

3. Simplify assembly graph

4. Detangle graph with long reads, mates, and other links
Why are genomes hard to assemble?

1. **Biological:**
   - (Very) High ploidy, heterozygosity, repeat content

2. **Sequencing:**
   - (Very) Large genomes, imperfect sequencing

3. **Computational:**
   - (Very) Large genomes, complex structure

4. **Accuracy:**
   - (Very) Hard to assess correctness
Ingredients for a good assembly

Coverage

High coverage is required
- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

Read Length

Reads & mates must be longer than the repeats
- Short reads will have false overlaps forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality

Errors obscure overlaps
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in de novo plant genome sequencing and assembly
Typical contig coverage

Imagine raindrops on a sidewalk
Balls in Bins 1x
Balls in Bins 2x
Balls in Bins 3x
Balls in Bins 4x
Balls in Bins 5x
Balls in Bins 6x
Balls in Bins 7x
Balls in Bins 8x
Coverage and Read Length

Idealized Lander-Waterman model
- Reads start at perfectly random positions
- Contig length is a function of coverage and read length
  - Short reads require much higher coverage to reach same expected contig length
- Need even high coverage for higher ploidy, sequencing errors, sequencing biases
  - Recommend 100x coverage

Assembly of Large Genomes using Second Generation Sequencing
Unitigging / Unipathning

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka “unitigs”, “unipaths”
**Repeats and Read Length**

- Explore the relationship between read length and contig N50 size
  - Idealized assembly of read lengths: 25, 35, 50, 100, 250, 500, 1000
  - Contig/Read length relationship depends on specific repeat composition

**Assembly Complexity of Prokaryotic Genomes using Short Reads.**
### Repetitive regions

<table>
<thead>
<tr>
<th>Repeat Type</th>
<th>Definition / Example</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-complexity DNA / Microsatellites</td>
<td>((b_1b_2...b_k)^N) where (1 \leq k \leq 6) CACACACACACACACACACACA</td>
<td>2%</td>
</tr>
<tr>
<td>SINEs (Short Interspersed Nuclear Elements)</td>
<td><em>Alu</em> sequence (~280 bp) Mariner elements (~80 bp)</td>
<td>13%</td>
</tr>
<tr>
<td>LINEs (Long Interspersed Nuclear Elements)</td>
<td>~500 – 5,000 bp</td>
<td>21%</td>
</tr>
<tr>
<td>LTR (long terminal repeat) retrotransposons</td>
<td><em>Ty1-copia</em>, <em>Ty3-gypsy</em>, <em>Pao-BEL</em> (~100 – 5,000 bp)</td>
<td>8%</td>
</tr>
<tr>
<td>Other DNA transposons</td>
<td></td>
<td>3%</td>
</tr>
<tr>
<td>Gene families &amp; segmental duplications</td>
<td></td>
<td>4%</td>
</tr>
</tbody>
</table>

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp
Error Correction with Quake

1. Count all “Q-mers” in reads
   - Fit coverage distribution to mixture model of errors and regular coverage
   - Automatically decide threshold for trusted k-mers

2. Correction Algorithm
   - Consider editing erroneous kmers into trusted kmers in decreasing likelihood
   - Includes quality values, nucleotide/nucleotide substitution rate

Quake: quality-aware detection and correction of sequencing reads.
Illumina Sequencing & Assembly

Quake Results
2x76bp @ 275bp
2x36bp @ 3400bp

Validated 51,243,281 88.5%
Corrected 2,763,380 4.8%
Trim Only 3,273,428 5.6%
Removed 606,251 1.0%

<table>
<thead>
<tr>
<th></th>
<th># ≥ 100bp</th>
<th>N50 (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffolds</td>
<td>2,340</td>
<td>253,186</td>
</tr>
<tr>
<td>Contigs</td>
<td>2,782</td>
<td>56,374</td>
</tr>
<tr>
<td>Unitigs</td>
<td>4,151</td>
<td>20,772</td>
</tr>
</tbody>
</table>
Outline

1. Ingredients for a good assembly

2. 2nd Generation Sequencing & Assembly
   1. Sacred Lotus
   2. Raspberry
   3. Wheat

3. 3rd Generation Sequence & Assembly
   1. Parrot
   2. Rice
Sacred Lotus Sequencing
*Nelumbo nucifera* Gaertn.

- Known for religious significance, herbal medicines, seed longevity, and water repellency
- Member of the Proteales, which lies outside of the core eudicots
  - Closest relatives are shrubs and trees belonging to the Proteaceae and Platanaceae
  - ~929Mbp Genome Size

Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.)
### Sacred Lotus Sequencing Approach

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Length</th>
<th>Fragment Length</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>100 bp</td>
<td>180 bp</td>
<td>33x</td>
</tr>
<tr>
<td></td>
<td>100 bp</td>
<td>500 bp</td>
<td>35x</td>
</tr>
<tr>
<td></td>
<td>35 bp</td>
<td>3,800 bp</td>
<td>6.4x</td>
</tr>
<tr>
<td></td>
<td>35 bp</td>
<td>8,000 bp</td>
<td>6.1x</td>
</tr>
<tr>
<td>454</td>
<td>*** 35 bp</td>
<td>20,000 bp</td>
<td>0.2x</td>
</tr>
</tbody>
</table>

**Flowchart:**
- **Raw Reads**
- **Merge pairs**
- **Correct Errors**
- **Build Unipaths**
- **Scaffold**
- **Finalize**
Sacred Lotus Assembly

Adding 20kbp mates improved scaffold N50 from 600kbp to 3.4Mbp

- Align 454 mates to draft assembly, extract the 35bp sequence from consensus
- Error corrects, remove duplicates

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Status</th>
<th>Number</th>
<th>N50 (kb)</th>
<th>Longest (kb)</th>
<th>size (Mb)</th>
<th>% cov</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>All</td>
<td>58409</td>
<td>38.8</td>
<td>286</td>
<td>707</td>
<td>76.1</td>
</tr>
<tr>
<td>Scaffold</td>
<td>All</td>
<td>3605</td>
<td>3435</td>
<td>14,300</td>
<td>804</td>
<td>86.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Annotation</th>
<th>number</th>
<th>Mean (bp)</th>
<th>Median (bp)</th>
<th>Length (Mb)</th>
<th>% genome</th>
<th>% GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>26,685</td>
<td>6562</td>
<td>3917</td>
<td>175</td>
<td>21.7</td>
<td>36</td>
</tr>
<tr>
<td>Exons</td>
<td>132,653</td>
<td>294</td>
<td>153</td>
<td>39</td>
<td>4.8</td>
<td>43</td>
</tr>
<tr>
<td>Introns</td>
<td>108,887</td>
<td>1249</td>
<td>283</td>
<td>136</td>
<td>16.9</td>
<td>34</td>
</tr>
<tr>
<td>TE</td>
<td>396,000</td>
<td>1111</td>
<td></td>
<td>440</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>Repeats</td>
<td>232,000</td>
<td>370</td>
<td></td>
<td>86</td>
<td>8.9</td>
<td></td>
</tr>
</tbody>
</table>
Raspberry Sequencing

*Rubus idaeus*

• Important food crop (~$1B / year in production). High amounts of fiber, vitamin C, manganese, and other nutrients

• Member of the Rosaceae family, along with other common fruits
  – Including apple, peach, and strawberry
  – ~350Mbp Genome Size

*The genome of the red raspberry (Rubus idaeus L.)*
Heterozygous Genomes

Raspberry effectively has 3 genomes
- 70% at full coverage
- 2x30% at half coverage

Basic assembly stats
- Scaffold N50: 17kbp
- Contig N50: 12kbp
Resolving the Heterozygosity

Chromosome 1  TATAATCAACCCGCTTGCCGATCTGATG
Chromosome 2  TATAATCAACCCACTTGCCGATCTGATG

- Exploring various approaches to identify and resolve the heterozygosity.
  - Improved scaffold N50 to more than 250kbp
  - Currently using genetic map to form larger linkage groups

De novo identification of “heterotigs” towards accurate and in-phase assembly of complex plant genomes

Wheat Sequencing

*Aegilops tauschii*

- One of the most important cereal crops in the world
- *A. tauschii* is one of the three ancestral species (DD) in modern bread wheat (*Triticum aestivum*)
  - Also looking to sequence other 2 species, and bread wheat
  - ~4.5Gbp Genome Size

In Collaboration with McCombie and Ware labs
## Wheat Sequencing & Assembly

<table>
<thead>
<tr>
<th>Technology</th>
<th>Read Length</th>
<th>Fragment Length</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina</td>
<td>100 bp</td>
<td>180 bp</td>
<td>69x</td>
</tr>
<tr>
<td></td>
<td>100 bp</td>
<td>300 bp</td>
<td>50x</td>
</tr>
<tr>
<td></td>
<td>35 bp</td>
<td>2,000 bp</td>
<td>6.6x</td>
</tr>
<tr>
<td></td>
<td>35 bp</td>
<td>5,000 bp</td>
<td>6.5x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Count</th>
<th>Max</th>
<th>N50</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffolds</td>
<td>97,313</td>
<td>2.76 Mbp</td>
<td>23,193</td>
<td>1.36 Gbp (30%)</td>
</tr>
<tr>
<td>Contigs</td>
<td>556,767</td>
<td>165 kbp</td>
<td>4,623</td>
<td>928 Mbp (20%)</td>
</tr>
</tbody>
</table>

- Poor coverage of the genome due to extreme repeat content
  - Had to downsample reads to fit into RAM
  - Randomly discard reads covered by kmers that occur more than 500 times

- Coverage may be sufficient for “gene-space”
Outline

1. Ingredients for a good assembly

2. 2\textsuperscript{nd} Generation Sequencing & Assembly
   1. Sacred Lotus
   2. Raspberry
   3. Wheat

3. 3\textsuperscript{rd} Generation Sequence & Assembly
   1. Parrot
   2. Rice
Ingredients for a good assembly

**Coverage**

- High coverage is required
  - Oversample the genome to ensure every base is sequenced with long overlaps between reads
  - Biased coverage will also fragment assembly

**Read Length**

- Reads & mates must be longer than the repeats
  - Short reads will have false overlaps forming hairball assembly graphs
  - With long enough reads, assemble entire chromosomes into contigs

**Quality**

- Errors obscure overlaps
  - Reads are assembled by finding kmers shared in pair of reads
  - High error rate requires very short seeds, increasing complexity and forming assembly hairballs

*Current challenges in de novo plant genome sequencing and assembly*
Hybrid Sequencing

**Illumina**
*Sequencing by Synthesis*

- High throughput (60Gbp/day)
- High accuracy (~99%)
- Short reads (~100bp)

**Pacific Biosciences**
*SMRT Sequencing*

- Lower throughput (600Mbp/day)
- Lower accuracy (~85%)
- Long reads (1-2kbp+)
SMRT Sequencing

Imaging of fluorescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).

SMRT Read Types

- **Standard sequencing**
  - Long inserts so that the polymerase can synthesize along a single strand

- **Circular consensus sequencing**
  - Short inserts, so polymerase can continue around the entire SMRTbell multiple times and generate multiple sub-reads from the same single molecule.

SMRT Sequencing Data

Yeast
(Pre-release Chemistry / 2010)

65 SMRT cells
734,151 reads after filtering
Mean: 642.3 +/- 587.3
Median: 553 Max: 8,495

Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch
Consistent quality across the entire read

- Uniform error rate, no apparent biases for GC/motifs
- Sampling artifacts at beginning and ends of alignments
Consensus Quality: Probability Review

Roll $n$ dice $\Rightarrow$ What is the probability that at least half are 6’s
(Consensus is wrong if at least half the bases are wrong)

<table>
<thead>
<tr>
<th>$n$</th>
<th>Min to Lose</th>
<th>Losing Events</th>
<th>$P$(Lose)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>$1/6$</td>
<td>16.7%</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$P(1 \text{ of } 2) + P(2 \text{ of } 2)$</td>
<td>30.5%</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>$P(2 \text{ of } 3) + P(3 \text{ of } 3)$</td>
<td>7.4%</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>$P(2 \text{ of } 4) + P(3 \text{ of } 4) + P(4 \text{ of } 4)$</td>
<td>13.2%</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>$P(3 \text{ of } 5) + P(4 \text{ of } 5) + P(5 \text{ of } 5)$</td>
<td>3.5%</td>
</tr>
</tbody>
</table>

$$n \quad \text{ceil}(n/2)$$

$$\sum_{i=\lceil n/2 \rceil}^{n} P(i \text{ of } n) = \sum_{i=\lceil n/2 \rceil}^{n} \binom{n}{i} (p)^i (1-p)^{n-i}$$
Consensus Accuracy and Coverage

Coverage can overcome random errors

- Dashed: error model from binomial sampling; solid: observed accuracy
- For same reason, CCS is extremely accurate when using 5+ subreads

\[ CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^i (1-e)^{n-i} \]
PacBio Error Correction

Correction Pipeline

1. Map short reads (SR) to long reads (LR)
2. Trim LR at coverage gaps
3. Compute consensus for each LR

Error corrected reads can be easily assembled, aligned

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina
1. Pre-overlap
   - Consistency checks

2. Trimming
   - Quality trimming & partial overlaps

3. Compute Overlaps
   - Find high quality overlaps

4. Error Correction
   - Evaluate difference in context of overlapping reads

5. Unitigging
   - Merge consistent reads

6. Scaffolding
   - Bundle mates, Order & Orient

7. Finalize Data
   - Build final consensus sequences

http://wgs-assembler.sf.net
SMRT-Assembly Results

Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case
*** Able to assemble entire microbial chromosomes into individual contigs ***
Improved Gene Reconstruction

200 kb

Illumina

454

454-PBcR

454-PBcR-Illumina

Assembly from Fragments

RefSeq Genes

FOXP2

GC Percent in 5-Base Windows

Repeating Elements by RepeatMasker

FOXP2 assembled on a single contig
Transcript Alignment

- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
  - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
  - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing

- New collaboration with Gingeras Lab looking at splicing in human
Internal Roadmap has made steady progress towards improving read length and throughput

Very recent improvements:
1. Improved enzyme:
   Maintains reactions longer
2. “Hot Start” technology:
   Maximize subreads
3. MagBead loading:
   Load longest fragments
PacBio Rice Sequencing

C1 Chemistry – Summer 2011
Median=639 Mean=824 Max=10,008

C2.5 Chemistry – Summer 2012
Median=2231 Mean=3290 Max=24,405
Preliminary Rice Assemblies

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Contig N50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina Fragments</td>
<td>3925</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>Illumina Mates</td>
<td>13696</td>
</tr>
<tr>
<td>50x 2x100bp @ 180</td>
<td></td>
</tr>
<tr>
<td>36x 2x50bp @ 2100</td>
<td></td>
</tr>
<tr>
<td>51x 2x50bp @ 4800</td>
<td></td>
</tr>
<tr>
<td>MiSeq Fragments</td>
<td>6444</td>
</tr>
<tr>
<td>23x 459bp</td>
<td></td>
</tr>
<tr>
<td>8x 2x251bp @ 450</td>
<td></td>
</tr>
<tr>
<td>PBeCR Reads</td>
<td>13600</td>
</tr>
<tr>
<td>6.3x 2146bp ** MiSeq for correction</td>
<td></td>
</tr>
<tr>
<td>PBeCR + Mates</td>
<td>In Progress</td>
</tr>
<tr>
<td>6.3x 2146bp ** MiSeq for correction</td>
<td></td>
</tr>
<tr>
<td>In collaboration with McCombie &amp; Ware labs @ CSHL</td>
<td></td>
</tr>
</tbody>
</table>
Long Read CNV Analysis

Aluminum tolerance in maize is important for drought resistance and protecting against nutrient deficiencies

• Segregating population localized a QTL on a BAC, but unable to genotype with Illumina sequencing because of high repeat content
• Long read PacBio sequencing revealed an additional copy of the ZnMATE1 membrane transporter and enabled assembly of the entire gene cluster

A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils
Maron, LG et al. (2012) Under review.
Why are crop genomes hard to assemble?

1. **Biological:**
   - (Very) High ploidy, heterozygosity, repeat content

2. **Sequencing:**
   - (Very) large genomes, imperfect sequencing

3. **Computational:**
   - (Very) Large genomes, complex structure

4. **Accuracy:**
   - (Very) Hard to assess correctness

With new biotechnologies and improved algorithms we can address these challenges

=> Cautiously optimistic
Acknowledgements

**Schatz Lab**
Giuseppe Narzisi
Shoshana Marcus
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
James Gurtowski
Alejandro Wences

**CSHL**
Hanan Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

**NBACC**
Adam Phillippy
Sergey Koren

**JHU/UMD**
Steven Salzberg
Mihai Pop
Ben Langmead
Cole Trapnell
Thank You!

http://schatzlab.cshl.edu/